Summary of ARC's ASDC DAAC Metadata Findings

August 16th, 2017

This report outlines the ARC team's metadata findings for the ASDC DAAC. The findings reported below are present in the majority of the collection level metadata records. Detailed reports for individual collection and granule level metadata records are provided separately. The detailed reports will be used for tracking the DAAC's progress towards working off the reported findings and metrics will be generated from the detailed reports. An explanation of the color coding found in the detailed reports is included in this report along with preliminary metrics.

1. **Collection Level Findings - ECHO 10**
    i. Currently the DataSetId is presented in an abbreviated format (for example: CAL_IIR_L1-Prov-V1-10). Because 'DataSetId' maps to 'EntryTitle' in CMR, the DataSetId should be a descriptive title of the dataset. Therefore, ARC recommends changing the information provided in the DataSetId element to a descriptive title of the dataset. If possible, this title should match the title of the dataset associated with the DOI.
    ii. For the dataset description element, many metadata records only provide the long name of the dataset as the description or only provide a description which is one sentence long. The 'Description' element is intended to provide a summary of the dataset, mimicking a journal abstract that is useful to the science community but also approachable for a first-time user of the data. ARC recommends replacing the short descriptions with the abstract provided on the dataset landing page.
    iii. A new DOI element has been added to the ECHO10 metadata schema (https://git.earthdata.nasa.gov/projects/EMFD/repos/echo-schemas/browse/schemas/10.0/Collection.xsd#226). DOI is a required element in CMR for NASA datasets. The DOI element is where the DOI string should be provided for all collection records. Here is a sample of how DOI should be provided in ECHO10 metadata:

```
<DOI>
        <DOI>10.5067/LIS/LIS-OTD/DATA308</DOI>
</DOI>
```

    iv. The 'ArchiveCenter' element is a GCMD controlled vocabulary field (https://gcmdservices.gsfc.nasa.gov/static/kms/providers/providers.csv). Please change "Atmospheric Science Data Center" to

"NASA/LARC/SD/ASDC." ASDC should contact GCMD directly if a change in the naming convention is desired.

v. ARC recommends providing the dataset citation in the 'CitationForExternalPublication' element. The citation provided on the dataset landing page should be leveraged whenever possible. Based on recent discussions at the metadata summits, it is likely that the 'CitationForExternalPublication' element in ECHO10 will be broken into several separate fields, rather than be a single block of text.

vi. The 'CollectionProgress' element is now a required element in UMM and is meant to describe the production status of the collection. Currently, there are three responses for the 'Collection Progress' element and those responses are chosen from a controlled vocabulary list. Valid responses are: PLANNED, IN WORK and COMPLETE. Note that 'Collection Progress' is the equivalent of the ECHO10 element called 'Collection State.' In the near future, the enumeration list for 'CollectionProgress' will be updated and expanded to include the following values:
    - PLANNED
    - ACTIVE
    - COMPLETE
    - NOT APPLICABLE

vii. A restriction flag is included in many of the ECHO10 records. Please consider including the 'RestrictionComment' element in order to explain the details of the restriction.

viii. Data format information is critical in determining data usability. ARC recommends including data format information whenever possible.

ix. The Temporal 'RangeDateTime' information should be as accurate as possible by matching the temporal information in the first and last granule. The 'BeginningDateTime' should match the start date and time of the first granule in the collection. Similarly, the 'EndingDateTime' should match the stop date and time of the final granule in the collection.

x. For ECHO10 records, "Contacts/Contact/Role" must be chosen from the following list: ARCHIVER, DISTRIBUTOR, ORIGINATOR, PROCESSOR

xi. For the ECHO10 element 'Contacts/Contact/OrganizationPhones/Phone/Type,' please change "Phone" to "Telephone" to precisely match UMM controlled vocabulary.

xii. The contact email included in the metadata (larc@eos.nasa.gov) differs from the email on ASDC's website (support-asdc@earthdata.nasa.gov). Please verify which email is most appropriate and make changes to the metadata if necessary.

xiii. Many records are missing science keywords. Science keywords are a searchable facet in CMR and are required. Keywords should be selected from the GCMD 'Earth Science and Earth Science Services' keyword list (https://gcmdservices.gsfc.nasa.gov/static/kms/sciencekeywords/sciencekeywords.csv?ed_wiki_keywords_page). ARC has recommended specific keywords for each record in the detailed reports.

xiv. The Platform Short Name, Long Name, and Type are all GCMD controlled vocabulary fields. Many ECHO10 metadata records have have been affected by updates to the GCMD Platforms/Sources vocabulary list (https://gcmdservices.gsfc.nasa.gov/static/kms/platforms/platforms.csv?ed_wiki_keywords_page). Please ensure these elements are updated to comply with GCMD Version 8.5 vocabulary.

xv. Sensors that are repeats of the instrument listed may be removed from the metadata. Otherwise, sensors should should match vocabulary found in the GCMD Instruments/Sensors list (https://gcmdservices.gsfc.nasa.gov/static/kms/instruments/instruments.csv?ed_wiki_keywords_page).

xvi. In addition to the 'Campaign/ShortName' element, which is required in CMR, ARC also recommends providing the 'Campaign/LongName' since the short name is often comprised of acronyms. Both the campaign short name and long name should come from the GCMD Projects vocabulary list (https://gcmdservices.gsfc.nasa.gov/static/kms/projects/projects.csv?ed_wiki_keywords_page).

xvii. Whenever feasible, the Online Access URL in the metadata should point as directly as possible to the described data. Pointing more directly to the appropriate data download folder eliminates confusion for the user and eases data accessibility.

xviii. Whenever possible, a link to the dataset landing page should be included for each collection level record. Ideally this link will leverage the DOI URL. More information on DOI landing pages and the information required for a dataset landing page can be found here: https://wiki.earthdata.nasa.gov/display/DOIsforEOSDIS/DOI+Landing+Page

xix. If possible, ARC recommends including supplemental links in the collection level metadata, especially the user's guide.

xx. ARC recommends providing a description for all Online Access URLs and all Online Resource URLs. Descriptions should be unique to each URL.

xxi. ARC recommends adding an Online Resource URL which links to ASDC OPeNDAP services for all relevant collections. Mime type values should be

provided for all services and should be selected from the values provided in the UMM-Common documentation. ARC and the CMR team can provide guidance on mime types as needed. ARC recommends listing "OPENDAP DATA" as the URL Type associated with OPeNDAP links.

xxii. ARC recommends providing information on the datum in the metadata, if possible.

2. Collection Level Findings - DIF 10

i. ARC recommends including dataset citation information in the "Dataset_Citation" elements whenever possible.

ii. If a DOI has been assigned to a dataset, ARC recommends the use of the DOI link in the 'Dataset_Citation/Online_Resource' element.

iii. 'Data_Set_Progress' is now a required element and is meant to describe the production status of the collection. There are three responses for the 'Data_Set_Progress' element and those responses are chosen from a controlled vocabulary list. Valid responses are: PLANNED, IN WORK and COMPLETE. As mentioned previously, this enumeration list will soon be updated and expanded to include the following values:

- PLANNED
- ACTIVE
- COMPLETE
- NOT APPLICABLE

iv. The 'Organization/Organization_Name/Short_Name' field is a GCMD controlled vocabulary field. Please change "ASDC" to "NASA/LARC/SD/ASDC". ASDC should contact GCMD directly if a change in the naming convention is desired. Please note that all collections from a DAAC should use the same naming convention for consistency.

v. The 'Organization/Organization_Name/Long_Name' field is a GCMD controlled vocabulary field. Please change "Atmospheric Science Data Center" to "Atmospheric Science Data Center, Science Directorate, Langley Research Center, NASA". ASDC should contact GCMD directly if a change in the naming convention is desired. Please note that all collections from a DAAC should use the same naming convention for consistency.

vi. ASDC currently provides some data access links via FTP. Access to data should be migrated from ftp to https to comply with EOSDIS policy. URS authentication should be implemented in front of data access.

vii. ARC recommends adding an Online Resource URL which links to ASDC OPeNDAP services for all relevant collections. Whenever possible, the OPenDAP URL should point as directly as possible to the described

collection's directory. Mime type values should be provided for all services and should be selected from the values provided in the UMM-Common documentation. ARC and the CMR team can provide guidance on mime types as needed. ARC recommends listing "OPENDAP DATA" as the URL Type associated with OPeNDAP links.

viii. All links and references to REVERB/ECHO should be removed from ASDC's metadata since Reverb will be retired in January 2018.

ix. ARC recommends that all URLs include a URL description ('Related_URL/ Description'). The description should be unique to each URL. If a new URL is added to the metadata, the URL should have a description.

x. For the 'Version_Description' element, because the description reads "Not provided," it can be removed from the metadata. Otherwise, please describe the version of the dataset in more detail.

xi. ARC recommends providing information on the datum in the metadata, if possible.

xii. Whenever possible, the 'Metadata_Dates/Data_Creation' and 'Metadata_Dates/Data_Last_Revision' element should be populated with the known dates. If not possible, ARC recommends "Not provided" be replaced with the most appropriate string from the following list: unknown, present, unbounded, future

3. Granule Level Findings

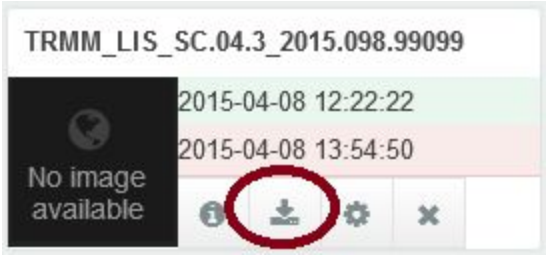
Findings common to all granule level records are presented first. Specific findings for ECHO 10 granule sheets 1 - 5 which correspond to ECHO 10 collection level sheets 1 -5 (ECS records) are presented second. Specific findings for ECHO 10 granule sheets 6 - 9 which correspond to DIF 10 collection level sheets 6 - 9 are presented last.

Common Findings Across All Granule Level Records

i. Measured parameter information is meant to "expose parameter level quality information about the granule." Therefore, the 'Measured Parameters/MeasuredParameter/ParameterName' element should only include the name of parameters found directly in the data files. It should be noted that this field will not be included in the UMM-G in the future. Parameter level metadata will be promoted to its own concept within the CMR via the unified variable model (UMM-Var). See UMM-G documentation for more information.

ii. ASDC currently provides some data access links via FTP. Access to data should be migrated from ftp to https to comply with EOSDIS policy. URS authentication should be implemented in front of data access.

iii. Whenever possible, an online access URL that points as directly to the referenced granule should be provided in the metadata. When this information is provided, there is a functionality in the Earthdata Search client that allows a user to directly download the file in the search client (after URS login).



iv. ARC recommends providing a description for all Online Access URLs and all Online Resource URLs. Descriptions should be unique to each URL.

v. Whenever possible, a link to the dataset landing page should be included for each granule level record. Ideally this link will leverage the DOI URL. More information on DOI landing pages and the information required for a dataset landing page can be found here: https://wiki.earthdata.nasa.gov/display/DOIsforEOSDIS/DOI+Landing+Page

vi. For OPeNDAP services, ARC recommends that a link to the specified granule be provided as an online resource URL in the granule level metadata. Mime type values should be provided for all services and should be selected from the values provided in the UMM-Common documentation. ARC and the CMR team can provide guidance on mime types as needed. ARC recommends listing "OPENDAP DATA" as the URL Type associated with OPeNDAP links. ARC has recommended in the detailed reports that the .HTML link to the OPeNDAP service be included. Note that the OPeNDAP link does not have to be formatted this way (.HTML) as long as the provided OPeNDAP link is consistent across the records.

vii. Please ensure all online resource URLs include an appropriate URL type. Valid values for URL type can be found in UMM-Common documentation as well as the UMM-Common schema:
URL type values:
https://git.earthdata.nasa.gov/projects/EMFD/repos/unified-metadata-model/browse/v1.9/umm-cmn-json-schema.json#1531

URL subtype values: https://git.earthdata.nasa.gov/projects/EMFD/repos/unified-metadata-model/browse/v1.9/umm-cmn-json-schema.json#1537

viii. ARC recommends including the campaign short name whenever possible. Note that campaign information is not inherited from the collection level record in CMR.

ECHO 10 Granule Sheets 1 - 5 (ECS Records)

ix. Currently the DataSetId is presented in an abbreviated format (for example: CAL_IIR_L1-Prov-V1-10). Because 'DataSetId' maps to 'EntryTitle' in CMR, the DataSetId should be a descriptive title of the dataset. Therefore, ARC recommends changing the information provided in the DataSetId element to a descriptive title of the dataset. This recommendation was made at the collection level.

x. Many granule records include a Restriction Flag. ARC recommends providing an explanation on the details of the restriction in the 'RestrictionComment' element since the restriction flag is only a number.

xi. For the 'Automatic Quality Flag Explanation' element, ARC recommends "QA flag explanation" be replaced with something more descriptive. Alternatively, ASDC can consider removing "QA flag explanation" from the metadata. Given that this element most likely populates as a part of an internal process, ASDC can judge the feasibility of removing or updating this information.

xii. Sensor short names should be removed if they are a repeat of the instrument short name. Sensor short names should also comply with the GCMD Instruments/Sensors list (https://gcmdservices.gsfc.nasa.gov/static/kms/instruments/instruments.csv?ed_wiki_keywords_page).

ECHO 10 Granule Sheets 6 - 9

xiii. ARC recommends that data format information be included in the granule level metadata. The data format should match the data format of the specified granule.

xiv. ARC recommends including the platform and instrument short names whenever possible. Note that in the Earthdata Search client/CMR, granule level records inherit the platform and instrument information from the collection level record.

4. Explanation of Color Coding Provided in Detailed Reports

Color (Metadata Element Names)	**Definition**
Cyan	Required field based on UMM-C
Light Purple	An optional primary element with required sub-elements based on UMM-C
Purple	A sub-element which is only required if any information is provided in the scope of the primary element based on UMM-C
White	Completely optional field
*	Any field with an asterisk is controlled by GCMD vocabulary

Color (Metadata Content)	**Definition**
Red	Correcting these issues should be of the highest priority
Yellow	Correcting these errors are highly recommended but are not required
Blue	Minor error/inconsistency; points out features noticed by the ARC Team which may help improve the robustness of the metadata but are not required to be addressed

5. Metrics

262 Collection level records checked

Collection Level	# Red fields	# Yellow fields	# Blue fields	Total # fields checked
	3,379	2,272	1,545	21,041
	16.1%	10.8%	7.3%	

248 Granule level records checked

Granule Level	# Red fields	# Yellow fields	# Blue fields	Total # fields checked
	670	670	1,682	8,998
	7.4%	7.4%	18.7%	

510 Total records checked (collection + granule)

Cumulative	# Red fields	# Yellow fields	# Blue fields	Total # fields checked
	4,049	2,942	3,227	30,039
	13.5%	9.8%	10.7%	

*Note - In collection level sheet 6, records MI3DRDNF (row 17) and MI3MRDNF (row 31) were reviewed but are now no longer in CMR. Metrics for these two records are included above.